applied sciences

MDPI

Article

An Intelligent Facial Expression Recognition System Using a Hybrid Deep Convolutional Neural Network for Multimedia Applications

Ahmed J. Obaid [1,2,*] and Hassanain K. Alrammahi [1]

[1] Faculty of Computer Science and Mathematics, University of Kufa, Najaf 54001, Iraq; hassanaink.alrammahi@uokufa.edu.iq
[2] Department of Computer Technical Engineering, Technical Engineering College, Al-Ayen University, Nasiriyah 64001, Iraq
* Correspondence: ahmedj.aljanaby@uokufa.edu.iq

Abstract: Recognizing facial expressions plays a crucial role in various multimedia applications, such as human–computer interactions and the functioning of autonomous vehicles. This paper introduces a hybrid feature extraction network model to bolster the discriminative capacity of emotional features for multimedia applications. The proposed model comprises a convolutional neural network (CNN) and deep belief network (DBN) series. First, a spatial CNN network processed static facial images, followed by a temporal CNN network. The CNNs were fine-tuned based on facial expression recognition (FER) datasets. A deep belief network (DBN) model was then applied to integrate the segment-level spatial and temporal features. Deep fusion networks were jointly used to learn spatiotemporal features for discrimination purposes. Due to its generalization capabilities, we used a multi-class support vector machine classifier to classify the seven basic emotions in the proposed model. The proposed model exhibited 98.14% recognition performance for the JaFFE database, 95.29% for the KDEF database, and 98.86% for the RaFD database. It is shown that the proposed method is effective for all three databases, compared with the previous schemes for JAFFE, KDEF, and RaFD databases.

Keywords: convolutional neural network; deep belief network; intelligent system; machine learning; human interaction

Citation: Obaid, A.J.; Alrammahi, H.K. An Intelligent Facial Expression Recognition System Using a Hybrid Deep Convolutional Neural Network for Multimedia Applications. *Appl. Sci.* **2023**, *13*, 12049. https://doi.org/10.3390/app132112049

Academic Editor: Douglas O'Shaughnessy

Received: 10 October 2023
Revised: 28 October 2023
Accepted: 4 November 2023
Published: 5 November 2023

1. Introduction

An individual's facial expressions (FEs) or countenance convey their psychological reactions and intentions in response to a social or personal event. These expressions convey non-verbal stealth messages. With technological advancements, human behavior can be understood through facial expression recognition (FER) [1]. To express emotions, humans use facial expressions as their primary nonverbal communication method [2]. Biometric authentication and nonverbal communication applications have recently drawn considerable attention to facial recognition. The movie industry has also conducted studies that predict emotions experienced during scenes. These works sought to identify a person's mood or emotion by analyzing facial expressions.

With landmarks in 2D and 3D [3], facial appearances [4], geometry [5], and 2D/3D spatiotemporal representations [6], facial models can be developed. Review papers provide comprehensive reviews [6–8]. It is also possible to categorize approaches based on images, deep learning, and model-based approaches. Engineered topographies such as HOGs [9], local binary pattern histograms (LBPHs) [10], and Gabor filters [11] are used in many image-based approaches.

The most recent works in this field focus on hand-engineered features, but various techniques have been developed [12–15]. In today's image processing and computer

vision applications, deep neural networks are the best choice due to the variety and size of datasets [16–18]. Conventional deep networks can easily handle spatial images [18]. Traditional feature extraction and classification schemes are computationally complex and difficult to use for achieving high recognition rates. A DNN based on convolutional coatings and residuals is proposed in this paper for recognizing facial emotions [19,20]. By learning the subtle features of each expression, the proposed model can distinguish them [21,22]. The proposed model presents a facial emotion classification report, along with the confusion matrix derived from the image dataset.

Afterwards, Section 2 presents a literature review with a deeper look at facial expression recognition based on deep learning. DBN and spatiotemporal CNN are integrated into Section 3 of the future model methodology. Section 4 grants the two sub-sections, the first sections present the three datasets, and the second section presents the untried outcomes and analysis. In the last section, we present the conclusion of the paper.

2. Literature Review

Human emotions are expressed through facial expressions in social communication. Three orthogonal planes are used to extract local binary pattern features (LBP-TOP) [23]. Based on computed histograms, the proposed LBP-TOP operator determines expression features from video sequences from three orthogonal planes. The author classified expressions using video sequences based on the extracted features of LBP-TOPs using a machine learning (ML) classifier. According to [24], fuzzy ARTMAP neural networks (FAMNNs) are used for VFER. In addition, particle swarm optimization (PSO) has been used to determine hyperparameters for the FAMNN network. A definite method [25] categorizes emotions based on their types, such as sadness, happiness, fear, and anger, according to the dimensional method [26].

SVM has also been used to categorize facial expressions [27]. Using 15 different feature points and their representations of neutral faces, the authors proposed a method measuring Euclidean distances between them. In the JAFFE dataset, 92% of the datasets are recognized, while in the CK dataset, 86.33% are recognized. These facial expression classification results demonstrate the effectiveness of SVMs in recognizing emotions. Also, SVMs have been employed for formalizing faces [28] and recognizing faces [29,30].

A large sample size is essential for developing algorithms for automatically recognizing facial expressions and related tasks. CK [31] and MMI [32] are three facial expression databases used to test and evaluate expression recognition algorithms. There are many databases where participants are asked to present certain facial expressions (e.g., frowns) rather than naturally occurring expressions (e.g., smiles). A spontaneous facial expression does not follow the same spatial or temporal pattern as a deliberately posed expression [33]. Over 90% of facial expressions can be detected by several algorithms. However, it is much harder to recognize spontaneous expressions [21,34]. A naturalistic setting is the best place for automatic FER. The working flow of FER methods [35], their integral and intermediate steps, and pattern structures are thoroughly analyzed and surveyed in this study in order to address these missing aspects. Furthermore, the limitations of existing FER surveys are discussed. A detailed analysis of FER datasets follows, followed by a discussion of the challenges and problems related to these datasets.

Deep-Learning-Based Face Recognition

During the training process, deep learning can auto-learn the new features based on stored information, thus minimizing the need to train the system repeatedly for new features. As deep learning algorithms do not require manual preprocessing, they can also handle large amounts of data. Recurrent neural networks (RNNs) and convolutional neural networks (CNNs) are two algorithms used in deep learning [36].

With RNN, relative dependencies with images are learned by recollecting information about past inputs, which is an advantage over CNN. It is widely used to combine RNNs and CNNs for image processing tasks, such as image recognition [37] and segmentation [38].

When input successions and hidden states are mapped to yields, a recurrent neural network (RNN) learns quick progression. In their paper, the authors proposed an improved method for representing spatial–temporal dependencies between two signals [39]. The CNN model is used in various fields, such as IOT [40], offloading [41], speech recognition [42], and traffic sign recognition [43].

A CNN and an RNN are combined in the HDCR-NN [25]. A facial expression classification system is adapted to it. Hybridizing convolutional neural networks [44] and recurrent neural networks [45] are automatically motivated by their wide acceptance of learning feature attributes. We used the Japanese Female Facial Expression (JAFFE) and Karolinska Directed Emotional Faces (KDEF) datasets to evaluate the proposed methodology.

Using cross-channel convolutional neural networks (CC-CNNs), the authors [46] propose a method for calculating VFER. The author of [47] proposes a method for VFER based on hierarchical bidirectional recurrent neural networks (PHRNNs). In their proposed framework (MSCNN), spatial information was extracted from still frames of an expression sequence using an MSCNN. Combining PHRNN and MSCNN further enhances VFER by extracting dynamic stills, parts and wholes, and geometry appearance information. An effective VFER method was demonstrated by combining spatiotemporal fusion and transfer learning. CNNs and RNNs are combined in the FER to incorporate audio and visual expressions.

An alternative to deep-learning-based methods used for recognizing facial emotions was proposed with a simple machine learning method [48]. Regional distortions are useful for analyzing facial expressions. On top of the convolutional neural network, they trained a manifold network to learn a variety of facial distortions. A set of low-level features has been considered for inferring human action [49] rather than only trying to extract facial features. An entropy-based method used for facial emotion classification was developed by [50]. An innovative method for recognizing facial expressions from videos was presented by [51]. Multiple feature descriptors have described face and motion information. To exploit complementary and discriminative distance metrics, we carried out the learning of multiple distance metrics. An ANN that learns and fuses the spatial–temporal features was proposed by [52]. The authors [50] proposed several preprocessing steps to recognize facial expressions.

3. Methodology

The proposed model comprises a convolutional neural network (CNN) and deep belief networks (DBNs). The spatial CNN network uses static facial images, followed by a temporal CNN network. After that, CNNs are fine-tuned based on facial expression recognition (FER) datasets. As a next step, the spatial and temporal characteristics of the segments are integrated using a deep belief network (DBN) model. Our hybrid deep learning model is shown in Figure 1. An optical flow network processes images generated between successive frame frames generated by a spatial CNN network, as shown in Figure 1. Using deep DBN models, these two CNN outputs are fused with the output of a fusion network. The CNN requires a fixed input data size, so each video sample is divided into several fixed-length segments. A fixed-length segment of L = 16 is created for each video sample.

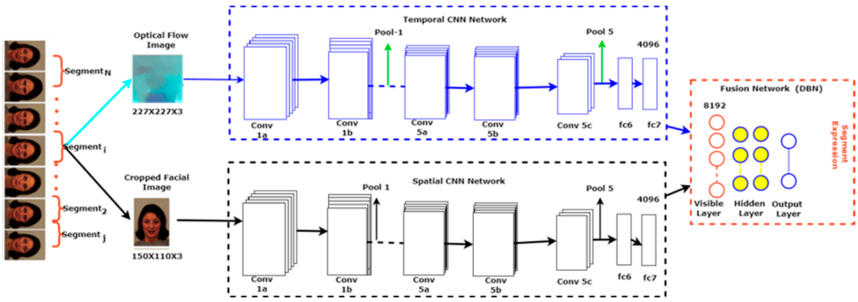


Figure 1. Facial expression model based on deep hybrid learning.

3.1. Image Resizing

Bilinear interpolation is used to resize the input image to a standard size. The real-life image size may vary, requiring the image to be resized. Displaying images on different devices requires image resizing. Resizing images is necessary for optimal display. Here, the image size is modified to attain a better outcome. The resized image is expressed as

$$R(F_i) = \{R(F_1), R(F_2), R(F_3), \ldots\ldots\ldots\ldots\ldots\ldots R(F_n)\} \quad (1)$$

where $R(F_i)$ indicates the resizing of the facial image and $R(F_n)$ indicates the resizing of the n-number of facial images. Finally, the facial images are resized into 227 × 227 × 3 for spatial CNN inputs. The first frame of a video segment L = 16 is discarded, and the remaining 15 frames are used as inputs for spatial CNNs.

3.2. Spatiotemporal Feature Learning with CNNs

The spatial and temporal CNNs are used same structure as VGG16 [53], which has five convolution layers (Conv1a-Conv1b, Conv2a-Conv2b, Conv3a-Conv3b-Conv3c-...-, Conv5a, and Conv5bConv5c), five max-pooling pools (Pool1, Pool2, and Pool3), and the FC layer consists of three layers (FC6, FC7, and FC8). The fc6 and fc7 represent 4096 units each, whereas fc8 represents data category labels. The VGG16 consists of 1000 image categories in the fc8 layer as shown in Figure 2.

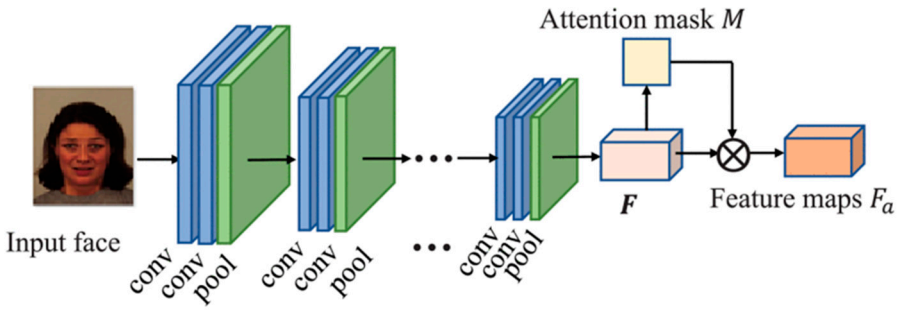


Figure 2. The CNN architecture for the proposed model.

Using CNNs, the VGG16 [53] learned spatiotemporal features from the video data of on-target facial expressions. Using pre-trained VGG16 parameters, both the temporal and spatial CNN networks are initialized. The facial expression category vectors in VGG16 are replaced with six new class label vectors. We then retrain each CNN stream individually using the backpropagation method. Backpropagation can solve the following minimization problem by updating the CNN network parameters:

$$\min_{w,\theta} \sum_{i=1}^{N} H(softmax(W \cdot \gamma(a_i; v)), y_i, \quad (2)$$

The network parameter v determines the weights of W, whether the CNN is a spatial CNN or temporal. Input data (a_i) are represented by $\gamma(v_i; v)$, which have 4096-D outputs of fc7, and the segment class label vector i^{th} is represented by y_i

$$H(v, y) = -\sum_{j=1}^{C} y_j \log(y_j), \quad (3)$$

In the fc7 layers of spatial and temporal CNNs, high-level features are represented as learned representations of input face images following training.

3.3. Spatiotemporal Fusion with DBNs

A deep DBN model was constructed by concatenating the outputs of the spatial and temporal CNNs in the fc7 layers [54]. The FER is represented using a deep DBN model based on a combination of discriminative features.

RBMs are bipartite graphs stacked in DBN models to form multilayered neural networks [55]. Two RBMs comprised two hidden layers and one visible layer. This illustrates the structure of a DBN by showing two RBMs with two hidden layers, and DBNs can be used to learn multilayer generative models using multiple RBMs. Multiple RBMs can be used to learn multilayer generative models. Thus, DBNs can be used to discover distribution properties and learn hierarchical feature representations.

The DBN fusion network is trained in two steps [56].

1. As a bottom-up method of unsupervised pretraining, greedy layer-wise training is used. The logarithm of the probability of derivatives is used to update the weights of each RBM model:

$$\Delta w = \varepsilon\left(< v_i h_j >_{data} - v_i h_j >_{model}\right) \tag{4}$$

The learning rate ε is represented by ε, while the data expectation is represented by <.>. A v_i indicates that a node is visible and a h_j indicates that a node is hidden.

2. Network parameters are updated through the supervised fine-tuning stage with backpropagation. Specifically, supervised fine-tuning can be achieved by comparing input data to the reconstructed data using the following loss function.

$$L(x, x') = ||x, x'||_2^2 \tag{5}$$

A reconstruction error $||\ ||_2^2$ is the L2-norm reconstruction error, where x and x' are the input data.

3.4. Classification

It is the classification component that completes the proposed AFER system. A supervised machine learning technique is used to accomplish this operation. A variety of algorithms can be used, including artificial neural networks (ANNs), k-NNs, decision trees, or deep learning techniques (such as CNNs) [57,58]. The author of [59] proposed the SVM classifier for this study. Several fields have shown this classifier's accuracy over the years. Datasets that are small and medium will be more suitable for this method. Recognizing unlabeled data will require the SVM classifier to be trained with labelled data. It is important to determine the ideal hyperplane in order to separate two datasets with two distinct classes during the learning phase of an SVM.

To separate the members of one class from those of another, a hyperplane is constructed using the input space transformed into a higher-dimensional feature space. Based on statistical learning theory, it is a useful classification method. The data can be separated using many hyperplanes. To prevent misclassifications, only one can attain the maximum margin when determining the type of data, and the new data that belong to. $S = \{x_i,\ d_i)|\, i = 1, 2, \ldots..N\}$ are usual for N training designs, with $x_i = (x_{i1},\ x_{i2}, \ldots\ldots\ldots,\ x_{in}) \in R^n$ being an input vector indicating a space in which to enter data and $d_i \in \{-1,\ 1\}$ indicating the class of the label x_i. The linear reparability of S is assumed. Optimally separating the hyperplane with maximum margins is imperative to generalizing a linear SVM well. In particular, the SVM locates the optimal hyperplane for the pair (w, b) by maximizing the margin $2/||w||$. Between $wx + b = 1$ and $wx + b = -1$, the perpendicular hyperplane of w points towards the separating plane and $b \in R$.

In this example, $(w_0,\ b_0)$ corresponds to a separating hyperplane for S that is optimal. An SVM using linear decision functions can be calculated using a decision function f

$$f(x) = sgn(w_0 x^t + b_0) \tag{6}$$

A sign function and matrix transpose operator are represented by $sgn(\cdot)$and t, respectively.

This introduces the following optimization problem when classifying non-separable data linearly with several slack variables ξ_i.

$$Minimize \frac{1}{2} W^T W + C \sum_{i=1}^{N} \xi_i$$

$$Subject\ to\ d_i(W^T x_i) \geq 1 - \xi_i,\ \xi_i \geq 0,\ i = 1, \ldots\ldots\ldots N. \tag{7}$$

It also acts as the regularization parameter for the SVM classifier, penalizing vectors incorrectly classified or within the margin.

A solution based on optimization $\overline{\alpha} = (\overline{\alpha_1}, \overline{\alpha_2}, \ldots..\overline{\alpha_N})$ can be used to calculate the pair $(\overline{w}, \overline{b})$. This formula can be used to rewrite the nonlinear SVM decision function f:

$$f(x) = sgn(w^{-t} x + \overline{b}) = sgn\left(\sum_{i=1}^{N} \overline{\alpha_1} d_i K(x_i, x) + \overline{b}\right) \tag{8}$$

In a transformed space, $K(\cdot, \cdot)$ is the kernel function used for fitting the maximum margin hyperplane. This paper uses the Gaussian radial basis function (GRBF) to specify kernel functions. For vector operations, the $K(x_i, x_j) = exp\left(\frac{(x_i - x_j)^2}{2y^2}\right)$ GRBF spread is given by y, while the L_2 norm is given by $||.||$.

A simplified version of the SVM is illustrated in Equation (9).

$$y_i = sign((w, x_i) + b) \tag{9}$$

A maximum margin hyperplane is represented by (w, x_i), a bias is represented by b, a feature vector is represented by $x_i \in \mathbb{R}^d$, and the labels are represented by $y_i \in \{\pm 1\}$. SVMs typically use kernels to handle nonlinear data, but we use linear SVMs. Furthermore, binary classifiers cannot distinguish between the six basic emotions of AFER. A multi-class SVM classifier is therefore formed by combining six binary SVM classifiers.

4. Experiments Analysis

The proposed model was evaluated based on the three publicly available datasets. It is typically necessary to preprocess an input for optimization purposes. First, 16 facial frames were extracted from the landmarks of the face. A face region was resized into 227 × 227 × 3 in order for the next processing block, namely feature extraction, to be applied.

4.1. Dataset

Three publicly available databases were used to evaluate the proposed method: the JAFFE [60], RaFD [61,62], and KDEF [63] databases. Face images and challenging conditions were included in these databases.

Japanese Female Facial Expression (JAFFE): These databases contain various face images and challenging conditions. A sequence of 2–4 samples was provided for each expression, and the images were 256 × 256 pixels in size. In Figure 3, sample data from the JAFFE database are presented. As shown in Table 1, each prototypical expression had several images in the databases.

Table 1. The used facial expression datasets.

Dataset	FE	AN	DI	HA	SA	SU	NE	Total	Resolution
JAFFE	30	30	28	29	30	30	30	207	256 × 256
KDEF	140	140	140	140	140	140	140	980	562 × 762
RaFD	67	67	67	67	67	67	67	469	681 × 1024

Figure 3. A sample facial expression sequence of the JAFFE dataset.

Radboud Faces Database (RaFD): 67 face models expressing basic emotions were represented in this database with 536 images. FACS experts trained all face models to express prototypical basic emotions. Additionally, all pictures were validated by FACS coders and 238 non-expert judges (N = 238) [60]. Twenty-seventy-three pictures were included in this study of 39 white adults expressing anger (AN), disgust (DI), neutrality (NE), fear (FE), sadness (SA), happiness (HA), and surprise (SU). A Caucasian face seemed to be more accurate with algorithms for facial expression analysis than a non-Caucasian face. White faces allow comparisons with previous validations of emotional facial expression categorization methods [63] and across different facial databases [64]. The sample figures of the RaFD datasets are presented in Figure 4.

Figure 4. A sample facial expression sequence of the RaFD dataset.

Karolinska Directed Emotional Faces (KDEF): The KDEF [63] database trained and tested our models for emotion recognition. Each individual in the corpus displayed seven different emotional expressions from five perspectives. The group comprised 70 individuals aged 20–30 (35 men and 35 women). The images were all taken in a controlled environment with fixed image coordinates for eye and mouth placements [63], as shown in Figure 5.

Figure 5. A sample facial expression sequence of the KDEF dataset.

4.2. Performance Metric

The accuracy, precision, recall, and F1 score of the proposed were assessed through cross-validation. By using the confusion matrix, the following matrices could be calculated: true positives (*TP*) are predictions that are positive and their actual data are positive; true negatives (*TN*) are predictions that are negative and their actual data are also negative; false positives (*FP*) are predictions that are positive but their actual results are negative; and false negatives (*FN*) are predictions that are negative but their actual results are positive. Equations (10)–(13) can be used to measure the accuracy, precision, recall, and *F*1-*score* [64].

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FN + FP)} \tag{10}$$

$$Recall = \frac{(TP)}{(TP + FN)} \tag{11}$$

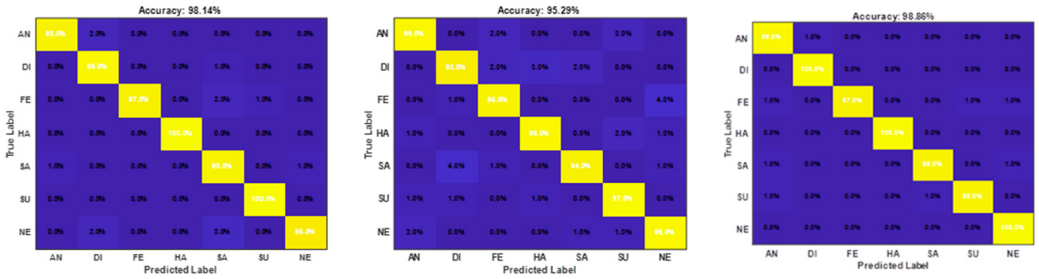
$$Precision = \frac{(TP)}{(TP + FP)} \tag{12}$$

$$F1 - Score = \frac{(2 \times P \times R)}{(P + R)} \tag{13}$$

4.3. Analysis and Discussion of Results

Three datasets were used in the proposed facial expression recognition model, and Figure 6 displays the confusion matrix. The JAFFEE model (Figure 6a) had the best overall results in terms of confusion matrix, followed by the KDEF and RaFD datasets (Figure 6c). According to the results, the "happy" class had the highest score, regardless of the model. This good prediction was due to two crucial factors: a large dataset and high variance. Training samples are important when highlighting the crucial features of a complex future. This can be seen by the fact that the neutral class overtook the happy class in the training samples. Although it had the largest share of the dataset, the "neutral" class failed to generalize as well as others. In the same way, anger, sadness, and surprise classes were compared. The classes shared the same proportion of data from the dataset, but their accuracy changed.

In all these observations, emotion variance played a significant role. The confusion matrix also demonstrates the similarity of emotion variation. It is easy to mistake fear for surprise and sad for neutrality. The "angry" and "sad" classes, for example, had a relatively low variance in mouth and eyebrow shape, which led to a significant number of images being classified as "neutral". The shape of the mouth did not change significantly, while the displacement of the eyebrows was hard to distinguish using the CNN model. Interestingly, the same results were not found vice versa, probably because there were so many training samples.



(**a**) (**b**) (**c**)

Figure 6. Confusion matrix for (**a**) JAFFE; (**b**) KDEF, and (**c**) RaFD datasets.

Three databases were used to run the proposed model. The proposed model produced intermediate feature maps, as shown in Table 2. Based on the confusion matrix given by the JAFFE, KDEF, and RaFD databases, Table 2 shows a classification report. Averaging the classification reports of each database 10-fold produced the classification report. A comparison was also made between the obtained results and those of existing state-of-the-art methods.

We compared the proposed model with the other approaches shown in Table 3 and Figure 7. The proposed model showed a 98.18% accuracy score for the JAFFE dataset. The last model is a fusion of several models corresponding to a method's accuracy. Existing models were outperformed by the proposed method. The proposed model combines deep belief networks (DBNs) with facial landmark coordinates to extract hybrid features superior to other deep-learning-based approaches. It is more accurate to classify expressions based on hybrid features. A spatiotemporal CNN with DBN can capture a relative correlation between facial landmarks associated with expressions based on landmark coordinates.

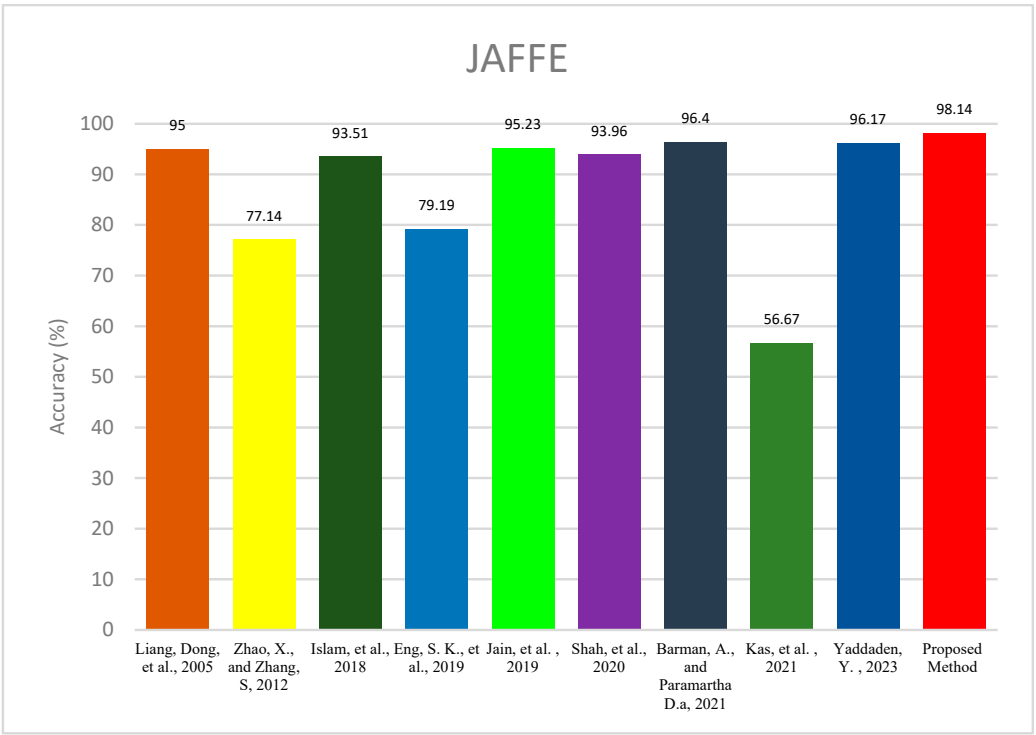


Figure 7. Graphical view of comparative analysis with state-of-the-art methods for the JAFFE dataset [57,65–72].

Table 2. Recognition rate (%) of the proposed model for the JAFFE, KDEF, and RaFD datasets.

Classes	JAFFE			KDEF			RaFD		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
NE	97.49	98.1	98.97	95.88	96.3	96.23	97.14	99.0	0.9832
AN	99.1	99.3	99.00	94.97	93.8	94.18	98.29	99.8	99.48
DI	97.73	97.2	97.43	94.11	95.4	94.97	95.97	96.9	96.90
FE	99.82	99.4	8.96	95.55	96.2	96.51	99.08	98.3	98.39
HA	96.39	97.9	97.11	95.80	94.2	95.12	98.95	99.2	99.56
SA	99.97	99.7	99.58	97.04	97.2	97.10	98.85	98.5	98.47
SU	95.79	95.6	95.71	95.27	95.8	95.83	99.50	99.1	98.84

Table 3. Comparing state-of-the-art methods for the JAFFE dataset.

Authors	Reference Number	Year of Study	Accuracy (%)
Liang, Dong, et al.	[67]	2005	95
Zhao, X., and Zhang, S	[68]	2012	77.14
Islam, et al.	[71]	2018	93.51
Eng, S.K., et al.	[70]	2019	79.19
Jain, et al.	[57]	2019	95.23
Shah, et al.	[65]	2020	93.96
Barman, A., and Paramartha D.a	[66]	2021	96.4
Kas, et al.	[69]	2021	56.67
Yaddaden, Y.	[72]	2023	96.17
Ahmed J. and Hassanain K.	Proposed model	2023	98.14

For the KDEF dataset, Table 4 compares the proposed model and the other FER models. A graphical representation of the comparison results is shown in Figure 8, and their accuracy scores are listed in Table 4. VGG19 [69] achieved an accuracy score of 76.73%. Another author's [72] model showed higher accuracy than the existing model, but the proposed models showed better results from all of these models. The proposed model on the EDEF database achieved a 95.29% accuracy rate. According to the authors [72], their model outperformed previous methods, proving its effectiveness.

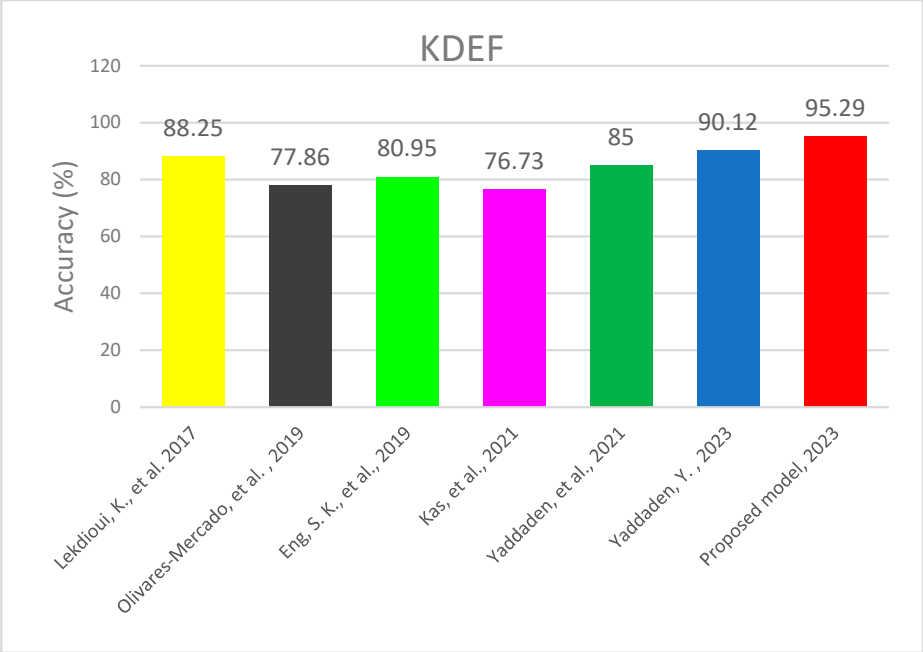


Figure 8. Graphical view of comparative analysis with state-of-the-art methods for the KDEF dataset [69,70,72–75].

Table 4. Comparing state-of-the-art methods for the KDEF dataset.

Authors	Reference Number	Year of Study	Accuracy
Lekdioui, K., et al.	[75]	2017	88.25
Olivares-Mercado, et al.	[73]	2019	77.86
Eng, S.K., et al.	[70]	2019	80.95
Kas, et al.	[69]	2021	76.73
Yaddaden, et al.	[74]	2021	85
Yaddaden, Y.	[72]	2023	90.12
Ahmed J. and Hassanain K.	Proposed model	2023	95.29

Using the RaFD dataset, the proposed method was compared to the other approaches. The correctness of 93.54% was achieved by combining CNNs, LBPs, and hog descriptors using another author's [74] model, as shown in Table 5 and Figure 9. Meanwhile, Ref. [76] presented an approach that can be used to recognize facial expressions in partial occlusion using Gabor filters and GLCM, which averages an accuracy score of 88.41%. According to ResNet50 [69], a person-independent FER framework was proposed that combines 49 landmarks' shapes and textural features to achieve 84.51% accuracy. The proposed method achieved 98.8% average accuracy compared to the other methods described above.

Table 5. Comparing state-of-the-art methods for the RaFD dataset.

Authors	Reference Number	Year of Study	Accuracy
Li, et al.	[76]	2015	88.41
Yaddaden, et al.	[74]	2021	93.54
Kas, et al.	[69]	2021	84.51
Yaddaden, Y.	[72]	2023	95.54
Ahmed J. and Hassanain K.	Proposed Method	2023	98.8

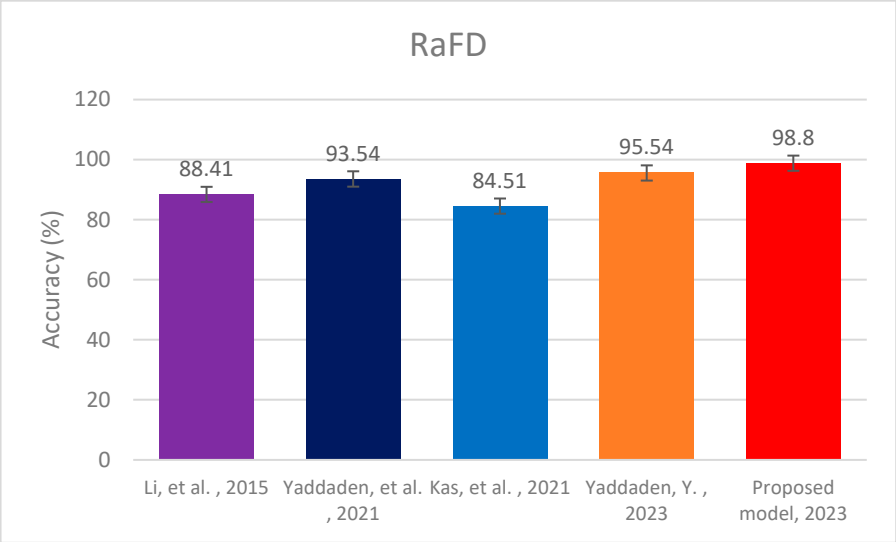


Figure 9. Graphical view of comparative analysis with state-of-the-art methods for the RaFD dataset [69,72,74,76].

Tables 3–5 show that our proposed model outperformed most previously studied models, although we used the same descriptors. There is a gap in accuracy among the three benchmark facial expression datasets. This improvement has been largely attributed to a multi-class SVM classifier and a CNN with DBN network models.

5. Conclusions

To effectively calculate the most important activities in effective calculations, people–computer interactions, machine vision, and consumer research into those facial expressions were processed. A person's facial expressions reveal their inner feelings and emotions, making them a form of nonverbal communication. This paper presented an efficient FER system using hybrid CNN with a DBN network. This model uses a dual-integrated CNN (CNN-DBN) that incorporates complementary knowledge from two CNN models trained independently on the FER datasets. Three publicly available FER datasets (JAFFEE, KDEF, and RaFD) were used to evaluate the proposed model. A comparative analysis of the results shows that the proposed model performs better than the existing models in terms of classification reports.

Author Contributions: Conceptualization, A.J.O.; methodology, H.K.A.; software, H.K.A.; validation, A.J.O.; formal analysis, A.J.O. investigation, A.J.O.; resources, H.K.A.; data curation, H.K.A.; writing—original draft preparation, H.K.A.; writing—review and editing, A.J.O.; visualization, H.K.A.; supervision, A.J.O.; project administration, A.J.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are openly available in [JAFFE, RaFD, KDEF] at [https://doi.org/10.1080/02699930903485076, https://doi.org/10.3390/app12178455], reference number [60–63].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shao, J.; Qian, Y. Three convolutional neural network models for facial expression recognition in the wild. *Neurocomputing* **2019**, *355*, 82–92. [CrossRef]
2. Joshi, D.; Datta, R.; Fedorovskaya, E.; Luong, Q.-T.; Wang, J.Z.; Li, J.; Luo, J. Aesthetics and emotions in images. *IEEE Signal Process. Mag.* **2011**, *28*, 94–115. [CrossRef]
3. Cootes, T.F.; Taylor, C.J.; Cooper, D.H.; Graham, J. Active Shape Models-Their Training and Application. *Comput. Vis. Image Underst.* **1995**, *61*, 38–59. [CrossRef]
4. Cootes, T.F.; Edwards, G.J.; Taylor, C.J. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 681–685. [CrossRef]
5. Kähler, K.; Haber, J.; Seidel, H.-P. Geometry-based muscle modeling for facial animation. In Proceedings of the Graphics Interface, Ottawa, ON, Canada, 7–9 June 2001; Volume 2001, pp. 37–46.
6. Fasel, B.; Luettin, J. Automatic facial expression analysis: A survey. *Pattern Recognit.* **2003**, *36*, 259–275. [CrossRef]
7. Li, S.; Deng, W. Deep Facial Expression Recognition: A Survey. *IEEE Trans. Affect. Comput.* **2022**, *13*, 1195–1215. [CrossRef]
8. Sandbach, G.; Zafeiriou, S.; Pantic, M.; Yin, L. Static and dynamic 3D facial expression recognition: A comprehensive survey. *Image Vis. Comput.* **2012**, *30*, 683–697. [CrossRef]
9. Carcagnì, P.; Del Coco, M.; Leo, M.; Distante, C. Facial expression recognition and histograms of oriented gradients: A comprehensive study. *SpringerPlus* **2015**, *4*, 1–25. [CrossRef]
10. Shan, C.; Gritti, T. Learning Discriminative LBP-Histogram Bins for Facial Expression Recognition. In Proceedings of the BMVC, Leeds, UK, 1–4 September 2008; pp. 1–10.
11. Lajevardi, S.M.; Lech, M. Averaged Gabor filter features for facial expression recognition. In Proceedings of the 2008 Digital Image Computing: Techniques and Applications, Canberra, Australia, 1–3 December 2008; pp. 71–76.
12. Kahou, S.E.; Froumenty, P.; Pal, C. Facial expression analysis based on high dimensional binary features. In Proceedings of the Computer Vision-ECCV 2014 Workshops, Zurich, Switzerland, 6–7,12 September 2014; Proceedings, Part II. Springer: Cham, Switzerland, 2015; pp. 135–147.
13. Shan, C.; Gong, S.; McOwan, P.W. Facial expression recognition based on local binary patterns: A comprehensive study. *Image Vis. Comput.* **2009**, *27*, 803–816. [CrossRef]
14. Rani, P.; Verma, S.; Yadav, S.P.; Rai, B.K.; Naruka, M.S.; Kumar, D. Simulation of the Lightweight Blockchain Technique Based on Privacy and Security for Healthcare Data for the Cloud System. *Int. J. E-Health Med. Commun.* **2022**, *13*, 1–15. [CrossRef]

15. Rani, P.; Singh, P.N.; Verma, S.; Ali, N.; Shukla, P.K.; Alhassan, M. An Implementation of Modified Blowfish Technique with Honey Bee Behavior Optimization for Load Balancing in Cloud System Environment. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 3365392. [CrossRef]
16. Kahou, S.E.; Bouthillier, X.; Lamblin, P.; Gulcehre, C.; Michalski, V.; Konda, K.; Jean, S.; Froumenty, P.; Dauphin, Y.; Boulanger-Lewandowski, N. Emonets: Multimodal deep learning approaches for emotion recognition in video. *J. Multimodal User Interfaces* **2016**, *10*, 99–111. [CrossRef]
17. Kalchbrenner, N.; Grefenstette, E.; Blunsom, P. A convolutional neural network for modelling sentences. *arXiv* **2014**, arXiv14042188.
18. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
19. Shamsolmoali, P.; Zareapoor, M.; Yang, J. Convolutional neural network in network (CNNiN): Hyperspectral image classification and dimensionality reduction. *IET Image Process.* **2019**, *13*, 246–253. [CrossRef]
20. Zareapoor, M.; Shamsolmoali, P.; Yang, J. Learning depth super-resolution by using multi-scale convolutional neural network. *J. Intell. Fuzzy Syst.* **2019**, *36*, 1773–1783. [CrossRef]
21. Sariyanidi, E.; Gunes, H.; Cavallaro, A. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 1113–1133. [CrossRef]
22. Sebe, N.; Lew, M.S.; Sun, Y.; Cohen, I.; Gevers, T.; Huang, T.S. Authentic facial expression analysis. *Image Vis. Comput.* **2007**, *25*, 1856–1863. [CrossRef]
23. Zhao, G.; Pietikainen, M. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 915–928. [CrossRef]
24. Gharavian, D.; Bejani, M.; Sheikhan, M. Audio-visual emotion recognition using FCBF feature selection method and particle swarm optimization for fuzzy ARTMAP neural networks. *Multimed. Tools Appl.* **2017**, *76*, 2331–2352. [CrossRef]
25. Jain, N.; Kumar, S.; Kumar, A.; Shamsolmoali, P.; Zareapoor, M. Hybrid deep neural networks for face emotion recognition. *Pattern Recognit. Lett.* **2018**, *115*, 101–106. [CrossRef]
26. Dellandrea, E.; Liu, N.; Chen, L. Classification of affective semantics in images based on discrete and dimensional models of emotions. In Proceedings of the 2010 International Workshop on Content Based Multimedia Indexing (CBMI), Grenoble, France, 23–25 June 2010; pp. 1–6.
27. Sohail, A.S.M.; Bhattacharya, P. Classifying facial expressions using level set method based lip contour detection and multi-class support vector machines. *Int. J. Pattern Recognit. Artif. Intell.* **2011**, *25*, 835–862. [CrossRef]
28. Khan, S.A.; Hussain, A.; Usman, M.; Nazir, M.; Riaz, N.; Mirza, A.M. Robust face recognition using computationally efficient features. *J. Intell. Fuzzy Syst.* **2014**, *27*, 3131–3143. [CrossRef]
29. Chelali, F.Z.; Djeradi, A. Face Recognition Using MLP and RBF Neural Network with Gabor and Discrete Wavelet Transform Characterization: A Comparative Study. *Math. Probl. Eng.* **2015**, *2015*, e523603. [CrossRef]
30. Ryu, S.-J.; Kirchner, M.; Lee, M.-J.; Lee, H.-K. Rotation invariant localization of duplicated image regions based on Zernike moments. *IEEE Trans. Inf. Forensics Secur.* **2013**, *8*, 1355–1370.
31. Kanade, T.; Cohn, J.F.; Tian, Y. Comprehensive database for facial expression analysis. In Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), Grenoble, France, 28–30 March 2000; pp. 46–53.
32. Pantic, M.; Valstar, M.; Rademaker, R.; Maat, L. Web-based database for facial expression analysis. In Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands, 6 July 2005. [CrossRef]
33. Wang, S.; Wu, C.; He, M.; Wang, J.; Ji, Q. Posed and spontaneous expression recognition through modeling their spatial patterns. *Mach. Vis. Appl.* **2015**, *26*, 219–231. [CrossRef]
34. Gunes, H.; Hung, H. Is automatic facial expression recognition of emotions coming to a dead end? The rise of the new kids on the block. *Image Vis. Comput.* **2016**, *55*, 6–8. [CrossRef]
35. Sajjad, M.; Ullah, F.U.M.; Ullah, M.; Christodoulou, G.; Cheikh, F.A.; Hijji, M.; Muhammad, K.; Rodrigues, J.J. A comprehensive survey on deep facial expression recognition: Challenges, applications, and future guidelines. *Alex. Eng. J.* **2023**, *68*, 817–840. [CrossRef]
36. Ansari, G.; Rani, P.; Kumar, V. A Novel Technique of Mixed Gas Identification Based on the Group Method of Data Handling (GMDH) on Time-Dependent MOX Gas Sensor Data. In *Proceedings of International Conference on Recent Trends in Computing*; Mahapatra, R.P., Peddoju, S.K., Roy, S., Parwekar, P., Eds.; Lecture Notes in Networks and Systems; Springer Nature: Singapore, 2023; Volume 600, pp. 641–654. ISBN 978-981-19882-4-0.
37. Visin, F.; Kastner, K.; Cho, K.; Matteucci, M.; Courville, A.; Bengio, Y. ReNet: A Recurrent Neural Network Based Alternative to Convolutional Networks. *arXiv* **2015**, arXiv150500393.
38. Sulong, G.B.; Wimmer, M.A. Image hiding by using spatial domain steganography. *Wasit J. Comput. Math. Sci.* **2023**, *2*, 39–45. [CrossRef]
39. Zhang, T.; Zheng, W.; Cui, Z.; Zong, Y.; Li, Y. Spatial–temporal recurrent neural network for emotion recognition. *IEEE Trans. Cybern.* **2018**, *49*, 839–847. [CrossRef]
40. Bhola, B.; Kumar, R.; Rani, P.; Sharma, R.; Mohammed, M.A.; Yadav, K.; Alotaibi, S.D.; Alkwai, L.M. Quality-enabled decentralized dynamic IoT platform with scalable resources integration. *IET Commun.* **2022**. [CrossRef]

41. Heidari, A.; Navimipour, N.J.; Jamali, M.A.J.; Akbarpour, S. A hybrid approach for latency and battery lifetime optimization in IoT devices through offloading and CNN learning. *Sustain. Comput. Inform. Syst.* **2023**, *39*, 100899. [CrossRef]
42. Alluhaidan, A.S.; Saidani, O.; Jahangir, R.; Nauman, M.A.; Neffati, O.S. Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network. *Appl. Sci.* **2023**, *13*, 4750. [CrossRef]
43. Triki, N.; Karray, M.; Ksantini, M. A real-time traffic sign recognition method using a new attention-based deep convolutional neural network for smart vehicles. *Appl. Sci.* **2023**, *13*, 4793. [CrossRef]
44. Zhou, W.; Jia, J. Training convolutional neural network for sketch recognition on large-scale dataset. *Int. Arab. J. Inf. Technol.* **2020**, *17*, 82–89. [CrossRef] [PubMed]
45. Zouari, R.; Boubaker, H.; Kherallah, M. RNN-LSTM Based Beta-Elliptic Model for Online Handwriting Script Identification. *Int. Arab. J. Inf. Technol.* **2018**, *15*, 532–539.
46. Barros, P.; Wermter, S. Developing crossmodal expression recognition based on a deep neural model. *Adapt. Behav.* **2016**, *24*, 373–396. [CrossRef]
47. Zhang, K.; Huang, Y.; Du, Y.; Wang, L. Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Trans. Image Process.* **2017**, *26*, 4193–4203. [CrossRef]
48. Jeong, M.; Ko, B.C. Driver's Facial Expression Recognition in Real-Time for Safe Driving. *Sensors* **2018**, *18*, 4270. [CrossRef] [PubMed]
49. Ullah, M.; Ullah, H.; Alseadonn, I.M. Human action recognition in videos using stable features. *Signal Image Process. Int. J.* **2017**, *8*, 1–10. [CrossRef]
50. Wang, S.-H.; Phillips, P.; Dong, Z.-C.; Zhang, Y.-D. Intelligent facial emotion recognition based on stationary wavelet entropy and Jaya algorithm. *Neurocomputing* **2018**, *272*, 668–676. [CrossRef]
51. Yan, H. Collaborative discriminative multi-metric learning for facial expression recognition in video. *Pattern Recognit.* **2018**, *75*, 33–40. [CrossRef]
52. Samadiani, N.; Huang, G.; Cai, B.; Luo, W.; Chi, C.-H.; Xiang, Y.; He, J. A Review on Automatic Facial Expression Recognition Systems Assisted by Multimodal Sensor Data. *Sensors* **2019**, *19*, 1863. [CrossRef]
53. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv14091556.
54. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507. [CrossRef] [PubMed]
55. Hinton, G.E. Training products of experts by minimizing contrastive divergence. *Neural Comput.* **2002**, *14*, 1771–1800. [CrossRef]
56. Hinton, G.E.; Osindero, S.; Teh, Y.-W. A fast learning algorithm for deep belief nets. *Neural Comput.* **2006**, *18*, 1527–1554. [CrossRef]
57. Jain, D.K.; Shamsolmoali, P.; Sehdev, P. Extended deep neural network for facial emotion recognition. *Pattern Recognit. Lett.* **2019**, *120*, 69–74. [CrossRef]
58. Yaddaden, Y.; Adda, M.; Bouzouane, A.; Gaboury, S.; Bouchard, B. User action and facial expression recognition for error detection system in an ambient assisted environment. *Expert Syst. Appl.* **2018**, *112*, 173–189. [CrossRef]
59. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
60. Lyons, M.J.; Akamatsu, S.; Kamachi, M.; Gyoba, J.; Budynek, J. The Japanese female facial expression (JAFFE) database. In Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 14–16 April 1998; pp. 14–16.
61. Langner, O.; Dotsch, R.; Bijlstra, G.; Wigboldus, D.H.; Hawk, S.T.; Van Knippenberg, A.D. Presentation and validation of the Radboud Faces Database. *Cogn. Emot.* **2010**, *24*, 1377–1388. [CrossRef]
62. Tsalera, E.; Papadakis, A.; Samarakou, M.; Voyiatzis, I. Feature Extraction with Handcrafted Methods and Convolutional Neural Networks for Facial Emotion Recognition. *Appl. Sci.* **2022**, *12*, 8455. [CrossRef]
63. Lundqvist, D.; Flykt, A.; Öhman, A. Karolinska directed emotional faces. *Cogn. Emot.* **1998**, 91.
64. Rani, P.; Sharma, R. Intelligent transportation system for internet of vehicles based vehicular networks for smart cities. *Comput. Electr. Eng.* **2023**, *105*, 108543. [CrossRef]
65. Shah, J.H.; Sharif, M.; Yasmin, M.; Fernandes, S.L. Facial expressions classification and false label reduction using LDA and threefold SVM. *Pattern Recognit. Lett.* **2020**, *139*, 166–173. [CrossRef]
66. Barman, A.; Dutta, P. Facial expression recognition using distance and shape signature features. *Pattern Recognit. Lett.* **2021**, *145*, 254–261. [CrossRef]
67. Liang, D.; Yang, J.; Zheng, Z.; Chang, Y. A facial expression recognition system based on supervised locally linear embedding. *Pattern Recognit. Lett.* **2005**, *26*, 2374–2389. [CrossRef]
68. Zhao, X.; Zhang, S. Facial expression recognition using local binary patterns and discriminant kernel locally linear embedding. *EURASIP J. Adv. Signal Process.* **2012**, *2012*, 20. [CrossRef]
69. Kas, M.; Ruichek, Y.; Messoussi, R. New framework for person-independent facial expression recognition combining textural and shape analysis through new feature extraction approach. *Inf. Sci.* **2021**, *549*, 200–220. [CrossRef]
70. Eng, S.K.; Ali, H.; Cheah, A.Y.; Chong, Y.F. Facial expression recognition in JAFFE and KDEF Datasets using histogram of oriented gradients and support vector machine. In *IOP Conference Series: Materials Science and Engineering*; IOP Publishing: Bristol, UK, 2019; Volume 705, p. 012031.
71. Islam, B.; Mahmud, F.; Hossain, A.; Goala, P.B.; Mia, M.S. A facial region segmentation based approach to recognize human emotion using fusion of HOG & LBP features and artificial neural network. In Proceedings of the 2018 4th International

Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT), Dhaka, Bangladesh, 13–15 September 2018; pp. 642–646.
72. Yaddaden, Y. An efficient facial expression recognition system with appearance-based fused descriptors. *Intell. Syst. Appl.* **2023**, *17*, 200166. [CrossRef]
73. Olivares-Mercado, J.; Toscano-Medina, K.; Sanchez-Perez, G.; Portillo-Portillo, J.; Perez-Meana, H.; Benitez-Garcia, G. Analysis of hand-crafted and learned feature extraction methods for real-time facial expression recognition. In Proceedings of the 2019 7th International Workshop on Biometrics and Forensics (IWBF), Cancun, Mexico, 2–3 May 2019; pp. 1–6.
74. Yaddaden, Y.; Adda, M.; Bouzouane, A. Facial expression recognition using locally linear embedding with lbp and hog descriptors. In Proceedings of the 2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-Being (IHSH), Boumerdes, Algeria, 9–10 February 2021; pp. 221–226.
75. Lekdioui, K.; Messoussi, R.; Ruichek, Y.; Chaabi, Y.; Touahni, R. Facial decomposition for expression recognition using texture/shape descriptors and SVM classifier. *Signal Process. Image Commun.* **2017**, *58*, 300–312. [CrossRef]
76. Li, R.; Liu, P.; Jia, K.; Wu, Q. Facial expression recognition under partial occlusion based on gabor filter and gray-level cooccurrence matrix. In Proceedings of the 2015 International Conference on Computational Intelligence and Communication Networks (CICN), Jabalpur, India, 12–14 December 2015; pp. 347–351.